Bioinformatics,
Volume 36, Issue 8, 15 April 2020, pp. 2626–2627
DOI:10.1093/bioinformatics/btz930

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Note

Data and text mining

VarGen: An R package for disease-associated variant discovery and annotation

Corentin Molitor[1], Matt Brember[1], Fady Mohareb[1*]

[1] The Bioinformatics Group, School of Water, Energy and Environment, Cranfield University, College Road, Bedford, MK43 0AL, UK.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Over the past decade, there has been an exponential increase in the amount of disease-related genomic data available in public databases. However, this high-quality information is spread across independent sources and researchers often need to access these separately. Hence, there is a growing need for tools that gather and compile this information in an easy and automated manner. Here we present "VarGen", an easy to use, customisable R package that fetches, annotates and rank variants related to diseases and genetic disorders, using a collection public databases (viz. OMIM, FANTOM5, GTEx and the GWAS catalog). This package is also capable of annotating these variants to identify the most impactful ones. We expect that this tool will benefit the research of variant-disease relationships.
Availability and implementation: VarGen is open-source and freely available via GitHub: https://github.com/MCorentin/VarGen. The software is implemented as an R package and is supported on Linux, MacOS and Windows.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Complex genetic diseases are often caused by the accumulation of a large number of low impacting variants rather than a single defective gene. With the current epidemics of complex non-transmittable diseases such as diabetes mellitus and obesity and the recent advances in sequencing technologies and genotyping, it is nowadays possible to gain comprehensive insights into the genetics behind these disorders. Moreover, there has been an exponential increase in the amount of high-quality information available in public databases, for example, the current build of dbSNP contains more than 660 million human RefSNP clusters (Sherry et al. 2001). Unfortunately, useful information is often scattered between independent sources, such as the Online Mendelian Inheritance in Man (OMIM), the Functional Annotation Of the Mammalian genome 5 (FANTOM5), the Genotype-Tissue Expression (GTEx) and Genome Wide Association Studies (GWAS). Each one of these databases provide useful and complementary information about the impact of variants on diseases but have to be accessed separately and sometimes are not based on the same version of the human genome. Some previous attempt to integrate SNP-related knowledge already exist [e.g. (Cao et al. 2017; Ferrero 2018; Pinero et al. 2017)], but these often lack completeness and/or the sensitivity required. Here we present VarGen, an easy-to-use R package for disease-associated variant discovery and annotation based on information integrated from different and complementary high-quality databases.

2 VarGen

VarGen implements a highly customisable workflow for retrieving and annotating SNPs using publicly available repositories (See Supplementary Figure 1). The workflow's entry point is typically a disease ID entered by the user. Alternatively, VarGen can retrieve causative SNPs based on a customised list of genes of interest. Genes related to the disease are first retrieved from the OMIM database (Amberger and Hamosh 2017), subsequently called the "OMIM genes". VarGen then retrieves variants situated directly on the OMIM genes, as well as variants present in their promoter regions using the FANTOM5 database. Integrating FANTOM5 data will allow the detection of non-coding variants, as there is more and more evidence of their non-negligible impact on diseases (Ward and Kellis 2012). Additionally, tissue-specific SNPs can also be retrieved using the Genotype-Tissue Expression eQTLs database

(Consortium et al. 2017). VarGen also accesses the GWAS catalogue to get variants associated with GWAS traits of interest (Buniello et al. 2019). VarGen, accesses these databases via BiomaRt (Smedley et al. 2009), Ensembl API and local files. The latter can be downloaded via the `vargen_install` function, making the installation straightforward. Moreover, all the positions are reported in hg38 coordinates. If a source still uses hg19, VarGen will lift-over the positions. The variants are then annotated according to their location, impact, clinical significance, and scored with the Combined Annotation Dependent Depletion (CADD) phred score (Rentzsch et al. 2019). Since VarGen typically outputs a large number of variants as a result of the comprehensive list of repositories queried, this annotation step is helpful to rank them and identify the most relevant.

VarPhen, a more specific alternative pipeline is also available within the package. VarPhen limits the variant output by retrieving the variants directly linked to a list of phenotypes in BiomaRt. The list of phenotypes is automatically obtained from keywords entered by the user as input (See Supplementary Figure 2).

In order to have an overview of the variants discovered by the pipeline, we developed a custom visualisation function, which displays the variants on each OMIM gene, grouped by according to their corresponding impact (See Supplementary Figure 3).

3 Benchmarking

VarGen was compared to two other similar tools: DisGeNET and VarFromPDB using the term “obesity” (OMIM: 601665) as a use case. The benchmarking script is available as supplementary data. Results of the benchmarking can be seen on Figure 1

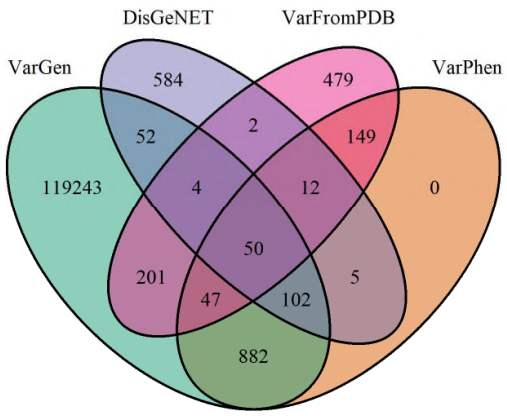


Figure 1: Venn diagram of the variants obtained with the different pipelines. Obesity (OMIM: 601665) was chosen as a use case. VarGen and VarPhen are the two alternative pipelines available in the package, focused on sensitivity and specifity respectively.

VarGen and VarPhen are sharing respectively 456 and 365 variants with DisGeNET and VarFromPDB. Moreover, 882 variants are shared only between VarGen and VarPhen, highlighting the higher sensitivity of both pipelines.

In total, DisGeNET and VarFromPDB are sharing 68 variants and only 2 of them are not discovered by either VarGen or VarPhen. Almost all of the 584 variants unique to DisGeNET are from literature mining and GwasDB which are not implemented in the other pipelines (see Supplementary Figure 4). Arguably, literature mining may introduce a large number of false positives and therefore was not included in our package. It was found that 408 variants out of the 479 variants unique to VarFromPDB are not associated directly with obesity but other phenotypes, such as Leptin dysfunction, Intellectual Disability, Bardet-Biedl syndrome; which therefore explains the limited overlap with the other tools (See Supplementary Figure 5 and Supplementary Table 1).

Some of the 119,243 unique variants from VarGen are potentially false positives or with little confirmed clinical evidence. It is possible to filter most of them using the phred score, source and clinical significance while keeping almost all the variants found by the other databases. (See Supplementary Figure 6). As some users will prefer a more specific approach, we provide an alternative pipeline, VarPhen, which gets only the most relevant variants.

Similarly, an additional benchmarking has been carried using Alzheimer’s disease (OMIM ID: 104300) as a use-case, where the results were comparable to these for obesity described above (See Supplementary Figures 7 to 9 and Supplementary Table 2).

4 Conclusions

VarGen is a flexible, well documented and easy to use R package for disease-related SNP discovery. The pipeline offers higher degree of sensitivity compared to other existing tools, notably because it uses databases often overlooked by other tools (e.g. FANTOM5). The output is a comprehensive list of annotated variants, ranked according to their phred score and clinical impact, which can also be visualised within a genome visualisation track.

Funding

Vargen was developed as part of the European Union’s Horizon 2020-funded project Nutrishield (GA 818110).

References

Amberger, J. S. and Hamosh, A. (2017), 'Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes', *Curr Protoc Bioinformatics,* 58, 1 2 1-1 2 12.

Buniello, A., et al. (2019), 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019', *Nucleic Acids Res,* 47 (D1), D1005-D12.

Cao, Zongfu, et al. (2017), 'VarfromPDB: An Automated and Integrated Tool to Mine Dis-ease-Gene-Variant Relations from the Public Databases and Literature', *Journal of Proteomics & Bioinformatics*.

GTEx Consortium (2017), 'Genetic effects on gene expression across human tissues', *Nature,* 550 (7675), 204-13.

Ferrero, E. (2018), 'Using regulatory genomics data to interpret the function of disease variants and prioritise genes from expression studies', *F1000Res,* 7, 121.

Pinero, J., et al. (2017), 'DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants', *Nucleic Acids Res,* 45 (D1), D833-D39.

Rentzsch, P., et al. (2019), 'CADD: predicting the deleteriousness of variants throughout the human genome', *Nucleic Acids Res,* 47 (D1), D886-D94.

Sherry, S. T., et al. (2001), 'dbSNP: the NCBI database of genetic variation', *Nucleic Acids Res,* 29 (1), 308-11.

Smedley, D., et al. (2009), 'BioMart--biological queries made easy', *BMC Genomics,* 10, 22.

Ward, L. D. and Kellis, M. (2012), 'Interpreting noncoding genetic variation in complex traits and human disease', *Nat Biotechnol,* 30 (11), 1095-106.

VarGen: An R package for disease-associated variant discovery and annotation.

Corentin Molitor[1], Matthew Brember [1], Fady Mohareb[1*]

[1] The Bioinformatics Group, School of Water, Energy and Environment, Cranfield University, College Road, Bedford, MK43 0AL, UK.

[*] To whom correspondence should be addressed.

Contact: f.mohareb@cranfield.ac.uk.

Supplementary Material S1: User Manual

VarGen

VarGen is an R package designed to get a list of variants related to a disease. It just need an OMIM morbid ID as input and optionally a list of tissues / gwas traits of interest to complete the results. You can use your own customised list of genes instead of an OMIM ID. VarGen is also capable of annotating the variants to help you rank and identify the most impactful ones.

All the coordinates are based on the hg38 version of the human genome.

VarGen is open-source and available on GitHub

Table of Contents

- Tips
  - How to plot the gwas variants
  - How to plot the omim variants

VarGen workflow

This pipeline is centred on the genes linked to the disease of interest in the Online Mendelian Inheritance in Man (subsequently called the "OMIM genes"). VarGen is designed as a discovery tool, if you want a more specific pipeline see "VarPhen" in "Alternative pipelines".

VarGen outputs variants from the following sources: - **OMIM:** Variants located directly on the "OMIM genes". - **FANTOM5:** Variants located on the enhancers / promoters of the "OMIM genes". - **GTEx:** Variants associated with a change in expression for the "OMIM genes", in certain tissues. Currently GTEx v7 and v8 are supported. - **GWAS catalog:** Variants associated with the phenotype of interest.

The variants are then annotated with myvariant.

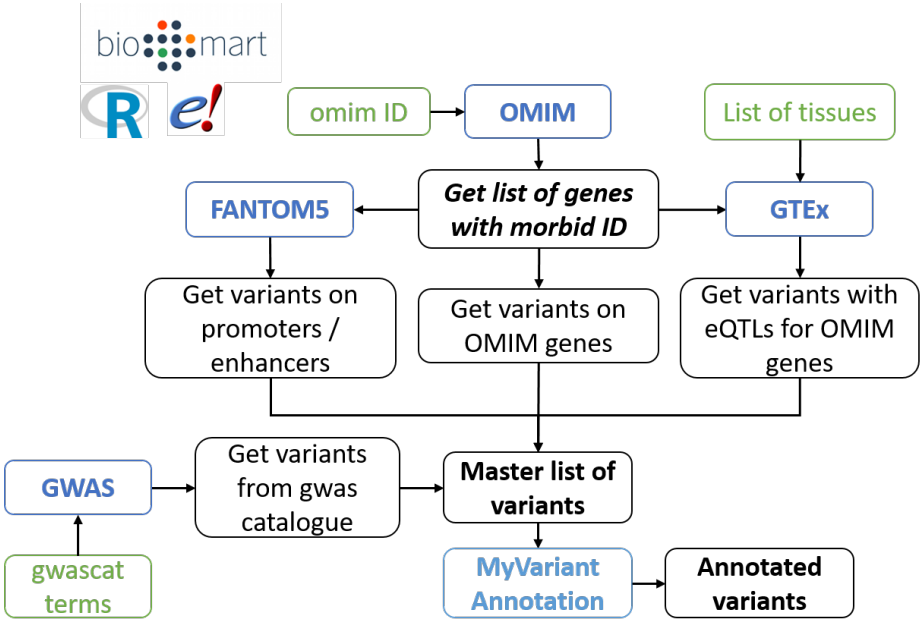


Figure 1: VarGen workflow

Installation

Dependencies

VarGen needs the following: - **R** (tested on version 3.6) - **An internet connection - The following R libraries:** (The number is the version tested during development)

`Bioconductor (3.9)`	`biomaRt (2.40.3)`	`gtools (3.8.1)`
`gwascat (2.16.0)`	`jsonlite (1.6)`	`GenomeInfoDb (1.20.0)`
`IRanges (2.18.1)`	`httr (1.4.1)`	`BiocGenerics (0.30.0)`
`stringr (1.4.0)`	`utils (3.6.0)`	`splitstackshape (1.4.8)`
`ggplot2 (3.2.1)`	`rtracklayer (1.44.2)`	`BiocManager (1.30.4)`
`R.utils (2.9.0)`	`myvariant (1.14.0)`	`GenomicRanges (1.36.0)`

- Optional R libraries (needed for the visualisation functions: *vargen_visualisation* and *plot_manhattan_gwas*):

`Gviz (1.28.1)`	`ggbio (>= 1.32.0)`	`grDevices (>= 3.6.0)`

To install the dependencies you can use the following command in R (it might take a while depending on your connection):

```
# If not already installed
install.packages("BiocManager")

BiocManager::install(c("biomaRt", "gtools", "GenomicRanges",
               "gwascat", "jsonlite",
                       "GenomeInfoDb", "IRanges", "httr",
               "BiocGenerics", "stringr",
                       "utils", "splitstackshape", "ggplot2", "rtracklayer",
                       "R.utils", "myvariant"), dependencies = TRUE)
```

note: "R.methodsS3" and "R.oo" will be installed as dependencies of "R.utils"

Install VarGen with devtools

The easiest way to get VarGen is to install it directly from R using "devtools":
To install devtools:

```
install.packages("devtools")
library(devtools)
install_github(repo = "MCorentin/VarGen", dependencies = TRUE)
library(VarGen)
```

Install VarGen from source

Alternatively you can clone the GitHub repository:

```
git clone https://github.com/MCorentin/VarGen
```

Then open R and install the package script from source:

```
library(utils)
install.packages("./VarGen/", repos = NULL, type = "source")
```

Preparing the input

Obtaining the local files

VarGen is fetching data from public databases, it needs the following files (they should all be in the same folder). The easiest way to get them is to use the *vargen_install* function from this package. The list of files obtained by the function is available below (with their approximate sizes and the links to download them manually). **note:** The installation can take a while, especially for the GTEx files.

```
vargen_install(install_dir = "./vargen_data",
        gtex_version = "v8", verbose = T)
```

Alternatively, they can be installed manually (the approximate size of every file is written in parenthesis next to the filename):

- **enhancer_tss_associations.bed** (~10 Mb), this is the enhancer to transcript start site association file from FANTOM5. (available at: http://enhancer.binf.ku.dk/presets/enhancer_tss_associations.bed)
- **hg19ToHg38.over.chain** (~500 kb), some databases are still using information from the human reference genome "hg19". VarGen will use this file to liftover the information to "hg38" (available at: http://hgdownload.cse.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz)
- **GTEx_Analysis_v8_eQTL** (~1.45 Gb), a folder containing the significant variant gene pairs from the Genotype-Tissues Expression database (GTEx) (available at: https://gtexportal.org/home/datasets). v7 and v8 are supported by VarGen.

Direct link for v7:

https://storage.googleapis.com/
gtex_analysis_v7/single_tissue_eqtl_data/
GTEx_Analysis_v7_eQTL.tar.gz

Direct link for v8:

https://storage.googleapis.com/
gtex_analysis_v8/single_tissue_qtl_data/
GTEx_Analysis_v8_eQTL.tar

- **gwas catalog file** (~90 Mb), this is an optional file. Depending on your connection, creating the lastest gwas catalog using *makeCurrentGwascat* function can take a long time. You can instead download a gwas catalog file from http://www.ebi.ac.uk/gwas/api/search/downloads/alternative and give it as input for the VarGen pipeline.

note: VarGen will use the name of the file to get the extract date, so it needs to be in the format: [filename]_r**YYYYY-MM-DD**.tsv

Getting the OMIM id

The OMIM morbid ID is the starting point of the pipeline, from it VarGen will get the genes associated to the disease. You can obtain it from the OMIM website https://omim.org/ or using the *list_omim_accessions* function. In our case we will use "obesity leanness included" (OMIM id: 601665).

```
gene_mart <- connect_to_gene_ensembl()
View(list_omim_accessions(gene_mart))

# You can search using list of keywords as well:
View(list_omim_accessions(gene_mart, c("alzheimer", "neurodegeneration")))
```

Getting the GTEx tissues

This database will be used to get tissue-specific variants that are affecting the expression of the genes related to the disease. You can obtain the list of files with the *list_gtex_tissues* function.

```
list_gtex_tissues(gtex_dir = "./vargen_data/GTEx_Analysis_v8_eQTL/")

#>       keywords                     filepaths
#> 1     Adipose_Subcutaneous         vargen_data/GTEx_Analysis_v8_eQTL/\
                     Adipose_Subcu.v8.signif_variant_gene_pairs.txt.gz
#> 2     Adipose_Visceral_Omentum     vargen_data/GTEx_Analysis_v8_eQTL/\
                                      Adipose_Vis_Omentum.v8.signif_variant_gene_pairs.txt.g
#> 3     Adrenal_Gland                vargen_data/GTEx_Analysis_v8_eQTL/\
                                      Adrenal_Gla.v8.signif_variant_gene_pairs.txt.gz
#> 4     Artery_Aorta                 vargen_data/GTEx_Analysis_v8_eQTL/\
                     Artery_Aor.v8.signif_variant_gene_pairs.txt.gz
#> ...
```

To select the GTEx tissues of interest you can use the *select_gtex_tissues* function, for our case we will use the adipose tissues (subcutaneous and visceral):

```
adipose_tissues <- select_gtex_tissues
    ("./vargen_data/GTEx_Analysis_v8_eQTL/", "adipose")
adipose_tissues

#> [1] "./vargen_data/GTEx_Analysis_v8_eQTL/\
Adipose_Subcutaneous.v8.signif_variant_gene_pairs.txt.gz"
#> [2] "./vargen_data/GTEx_Analysis_v8_eQTL/\
Adipose_Visceral_Omentum.v8.signif_variant_gene_pairs.txt.gz"
```

Getting the gwas traits

The gwas traits will be used to get associated variants in the gwas catalog (https://www.ebi.ac.uk/gwas/). You can search the available gwas traits by keyword with the *list_gwas_traits* function. Here we are going to use a local gwas catalog file (gwas_catalog_v1.0.2-associations_e96_r2019-08-24.tsv).

```
obesity_traits <- list_gwas_traits("obesity", "./vargen_data")
obesity_traits
```

```
#>  [1] "Obesity (extreme)"
#>  [2] "Obesity-related traits"
#>  [3] "Obesity"
#>  [4] "Obesity (early onset extreme)"
#>  [5] "Obesity in adult survivors of childhood
    cancer exposed to cranial radiation"
#>  [6] "Obesity in adult survivors of childhood
    cancer not exposed to cranial radiation"
#>  [7] "Bilirubin levels in extreme obesity"
#>  [8] "Type 2 diabetes (young onset) and obesity"
#>  [9] "Obesity and osteoporosis"
#> [10] "Hepatic lipid content in extreme obesity"
#> [11] "Hyperinsulinemia in obesity"
```

How to use VarGen

Launching the pipeline

VarGen main entry point is the *vargen_pipeline* function.

The mandatory inputs are: - The folder with the installed files, "./vargen_data/" in our case (see "Obtaining the local files") - A OMIM morbid id, "601665" in our case (see "Getting the OMIM id"). - An output directory, where information about the variants and the genes will be written. The default is the current directory "./".

The optional inputs are: - A fantom5 correlation threshold, the higher it is, the stricter you are about association between genes and enhancers. The default is 0.25 - A list of GTEx tissues (see "Getting the GTEx tissues") - A list of gwas traits (see "Getting the gwas traits")

Now we can launch the pipeline with all the input data:

```
adipose_tissues <- select_gtex_tissues("./vargen_data/GTEx_Analysis_v8_eQTL/",
                                       "adipose")

obesity_traits <- list_gwas_traits("obesity", "./vargen_data/")

obesity_variants <- vargen_pipeline(vargen_dir = "./vargen_data/",
```

```
                                    omim_morbid = "601665",
                                    gtex_tissues = adipose_tissues,
                                    gwas_traits = obesity_traits,
                                    verbose = T)
```

You will obtain a data.frame with the list of variants, their position on GRCh38, the ensembl id and hgnc gene symbol of the gene associated with the variant and the source (omim, fantom5, gtex or gwas).

```
head(obesity_variants)
```

```
#  chr     pos         rsid           ensembl_gene_id    hgnc_symbol  source
#> chr2    25160855    rs777983882    ENSG00000115138    POMC         omim
#> chr2    25160866    rs1480805741   ENSG00000115138    POMC         omim
#> chr2    25160871    rs1245939527   ENSG00000115138    POMC         omim
#> chr2    25160872    rs1219237056   ENSG00000115138    POMC         omim
#> chr2    25160877    rs1453226041   ENSG00000115138    POMC         omim
#> chr2    25160879    rs566456581    ENSG00000115138    POMC         omim
```

Annotating the variants

This pipeline is designed as a discovery analysis, to identify potential new variants, **you should not expect every variants from the pipeline to have an effect on the phenotype**. Annotating the variants will help you defining which variants to keep or discard. To annotate the variants you can use the *annotate_variants* function with the list of rsids obtained with the vargen pipeline. This uses myvariant.info to annotate the variants and may take some time depending on your internet connection.

The annotation contains: - CADD Phred score: ranging from 1 to 99, based on the rank of each variant relative to all possible 8.6 billion substitutions in the human reference genome. A higher value means a more deleterious variant. (above 10 means the variant is in the top 10%, above 20 in the top 1%). - fathmm-xf score: between 0 and 1, a higher value means a more deleterious variant. (more confidence closer to 0 or 1) - fathmm-xf prediction: "D" (DAMAGING) if score > 0.5 or "N" (NEUTRAL) otherwise. - **Annotation type**: information about the variant location (eg: coding, non-coding, regulatory region...) - **Consequence**: gives more information on the functionnal effect (eg: REGULATORY, DOWNSTREAM, STOP_GAINED, SPLICE-SITE...) - ClinVar clinical significance: standard to report the clinical significance of certain variants (eg: "benign", "pathogenic", "drug response" etc....). - snpEff impact: assessment of the putative impact of the variant (HIGH, MODERATE, MODIFIER or LOW).

/! Due to the different transcripts for a same gene, some variants will appear more than once with a different annotation, this is expected. You can check the variant position on the different isoforms with the *vargen_visualisation* function.

```
obesity_annotation <- annotate_variants(obesity_variants$rsid, verbose = T)

# Merging the output of the annotation with VarGen output, using the "rsid" column
obesity_ann <- merge(obesity_variants, obesity_annotation)

# We advise you to write the variants in a file,
#so you will not have to run the pipeline again.
write.table(x = obesity_ann, quote = FALSE, row.names = FALSE,
            file = "./OMIM_601665/vargen_variants_annotated.tsv", sep = "\t")
```

Output after annotation:

rsid	chr	pos	ensembl_gene_id	hgnc_symbol	source	cadd_phred	fathmm_xf_score	fathmm_xf_pred	annot_type	consequence	clinical_significance	snpeff_ann
rs17817449	chr16	53779455	ENSG00000140718	FTO	gwas	NA	NA	NA				MODIFIER
rs17817469	chr3	12368850	ENSG00000132170	PPARG	omim	1.928	NA	NA	Transcript	INTRONIC		MODIFIER;MODIFIER;MODIFIER;MODIFIER
rs17847043	chr6	131872658	ENSG00000197594	ENPP1	omim	0.499	NA	NA	Transcript	INTRONIC		MODIFIER
rs17847044	chr6	131873208	ENSG00000197594	ENPP1	omim	NA	NA	NA	NA	NA		MODIFIER
rs17847045	chr6	131878802	ENSG00000197594	ENPP1	omim	3.230	NA	NA	Transcript	INTRONIC		MODIFIER
rs17847045	chr6	131878802	ENSG00000197594	ENPP1	omim	NA	NA	NA	NA	NA		MODIFIER
rs17847047	chr6	131890297	ENSG00000197594	ENPP1	omim	0.987	NA	NA	Transcript	INTRONIC		
rs17847050	chr6	131860393	ENSG00000197594	ENPP1	omim	25.900	0.944065	D	CodingTranscript	NON_SYNONYMOUS	Uncertain significance;...	MODERATE
rs17848365	chr11	74008957	ENSG00000175564	UCP3	omim	NA	NA	NA				MODIFIER;MODIFIER;MODIFIER
rs17848366	chr11	74006459	ENSG00000175564	UCP3	omim	9.549	NA	NA	Transcript	INTRONIC		MODIFIER;MODIFIER
rs17848367	chr11	74006351	ENSG00000175564	UCP3	omim	9.179	0.078845	N	CodingTranscript	NON_SYNONYMOUS		MODERATE;MODERATE
rs17848368	chr11	74006298	ENSG00000175564	UCP3	omim	33.000	0.725808	D	CodingTranscript	NON_SYNONYMOUS	Pathogenic	MODERATE;MODERATE
rs17848372	chr11	74005915	ENSG00000175564	UCP3	omim	35.000	0.927409	D	CodingTranscript	NON_SYNONYMOUS		MODERATE;MODERATE
rs17851080	chr1	30877089	ENSG00000162512	SDC3	omim	0.020	NA	NA	CodingTranscript	SYNONYMOUS		LOW
rs17854409	chr20	62860142	ENSG00000101190	TCFL5	gwas	24.900	0.148471	N	CodingTranscript	NON_SYNONYMOUS		MODERATE;MODERATE
rs17854409	chr20	62860142	ENSG00000101190	TCFL5	gwas	19.450	0.109146	N	CodingTranscript	NON_SYNONYMOUS		MODERATE;MODERATE
rs1795708	chr12	58356897		RPL21P103 - ...	gwas	NA	NA	NA				MODIFIER
rs1797878	chr3	12417857	ENSG00000132170	PPARG	omim	NA	NA	NA	NA	NA		MODIFIER;MODIFIER;MODIFIER;MODIFIER
rs1797879	chr3	12417860	ENSG00000132170	PPARG	omim	NA	NA	NA	NA	NA		MODIFIER;MODIFIER;MODIFIER;MODIFIER
rs1797895	chr3	12406315	ENSG00000132170	PPARG	omim	3.303	NA	NA	RegulatoryFeature;Transcript	REGULATORY;INTRONIC		MODIFIER;MODIFIER;MODIFIER;MODIFIER
rs1797912	chr3	12428740	ENSG00000132170	PPARG	omim	2.285	NA	NA	Transcript	INTRONIC		MODIFIER;MODIFIER;MODIFIER;MODIFIER

Filtering the variants

As VarGen is outputting a lot of variants, you may want to filter the results. The filtering strategy is dependent on your preferences. VarGen outputs an R data.frame containing information about each variant. You can focus on clinically significant variants ("pathogenic", "likely pathogenic" etc...), on variants with a high phred score etc...

You can even combine different filtering, for example we found that keeping all the variants from the gwas catalog and with clinical significance while removing the variants with a cadd phred score lower than 10 was removing a lot of potentially false positive results:

```
vargen_phred_10 <- obesity_ann [obesity_ann$cadd_phred > 10,]
vargen_phred_10 <- vargen_phred_10[!is.na(vargen_phred_10$cadd_phred),]

vargen_clinVar <- obesity_ann [obesity_ann$clinical_significance != "",]

vargen_gwas <- obesity_ann[obesity_ann$source == "gwas",]

#Concatenating the different filtering
obesity_filtered <- rbind(vargen_phred_10, vargen_clinVar, vargen_gwas)
```

Alternative pipelines

VarPhen

A more specific, alternative pipeline is available as part of this package, called "VarPhen", it outputs a smaller list of variants, but directly related to the disease of interest. It relies on biomaRt to link variants to phenotypes.

note: You will need to specify the columns to merge the results from VarPhen ("redsnp_id") and the annotation ("rsid"). cf: the example below.

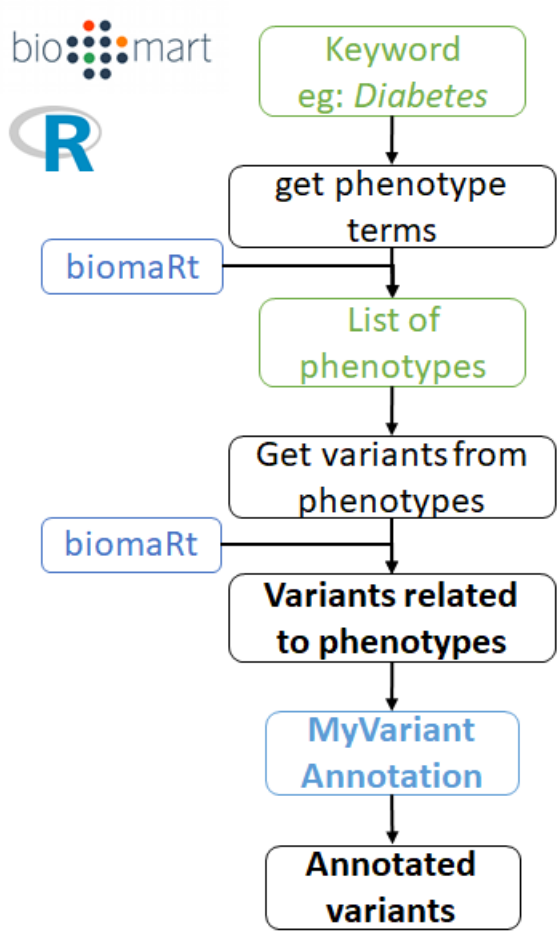


Figure 2: VarPhen workflow

Example with obesity:

```
# First connect to snp ensembl
```

```
snp_mart <- connect_to_snp_ensembl()

obesity_phens <- get_phenotype_terms(keyword = "obesity",
                                     snp_mart = snp_mart)

obesity_varphen <- get_variants_from_phenotypes(phenotypes = obesity_phens,
                                                snp_mart = snp_mart)

obesity_varphen_annotation <- annotate_variants(obesity_varphen$refsnp_id)

# /!\ contrary to vargen, here you need to specify which columns to merge
# "refsnp_id" for varphen, and "rsid" for the annotation.
obesity_varphen_annotated <- merge(obesity_varphen,
                                   obesity_varphen_annotation,
                                   by.x = "refsnp_id", by.y = "rsid")
```

Output example:

chr_name	chrom_start	refsnp_id	refsnp_source	associated_variant_risk_allele	phenotype_description	clinical_significance	validated
chr19	39830443	rs367643250	dbSNP	A	ABDOMINAL OBESITY-METABOLIC SYNDROME 3	pathogenic	Cited,ESPExAC,Frequency,gnomAD,Phenotype_or_Diseas...
CHR_HG2021_PATCH	39830443	rs367643250	dbSNP	A	ABDOMINAL OBESITY-METABOLIC SYNDROME 3	pathogenic	Cited,ESPExAC,Frequency,gnomAD,Phenotype_or_Diseas...
chr19	39830478	rs587777380	dbSNP	G	ABDOMINAL OBESITY-METABOLIC SYNDROME 3	pathogenic	Cited,Phenotype_or_Disease
CHR_HG2021_PATCH	39830478	rs587777380	dbSNP	G	ABDOMINAL OBESITY-METABOLIC SYNDROME 3	pathogenic	Cited,Phenotype_or_Disease
chr1	27549343	rs777736953	dbSNP	A	ABDOMINAL OBESITY-METABOLIC SYNDROME 3	likely pathogenic	ExAC,Frequency,gnomAD,Phenotype_or_Disease
chr2	233764663	rs4148325	dbSNP		Bilirubin levels in extreme obesity		1000Genomes,Cited,Frequency,gnomAD,HapMap,Pheno...
chr6	78970876	rs200788163	dbSNP	A	DEVELOPMENTAL DELAY INTELLECTUAL DISABILITY OBESIT...	uncertain significance,pathogenic	1000Genomes,ExAC,Frequency,gnomAD,Phenotype_or_D...
chr6	79060680	rs768324201	dbSNP	T	DEVELOPMENTAL DELAY INTELLECTUAL DISABILITY OBESIT...	pathogenic	Phenotype_or_Disease
chr6	79077904	rs878854420	dbSNP	G	DEVELOPMENTAL DELAY INTELLECTUAL DISABILITY OBESIT...	likely pathogenic,pathogenic	Cited,Phenotype_or_Disease
chr6	79025986	rs878854421	dbSNP		DEVELOPMENTAL DELAY INTELLECTUAL DISABILITY OBESIT...	likely pathogenic,pathogenic	Cited,Phenotype_or_Disease
chr6	79060709	rs1554212744	ClinVar		DEVELOPMENTAL DELAY INTELLECTUAL DISABILITY OBESIT...	pathogenic	Phenotype_or_Disease
chr6	78961775	rs1554196907	ClinVar	A	DEVELOPMENTAL DELAY INTELLECTUAL DISABILITY OBESIT...	pathogenic	Phenotype_or_Disease
chr7	45195521	rs1294908	dbSNP		Hepatic lipid content in extreme obesity		1000Genomes,Cited,Frequency,gnomAD,HapMap,Pheno...
chr22	43932850	rs4823173	dbSNP		Hepatic lipid content in extreme obesity		1000Genomes,Cited,ESPExAC,Frequency,gnomAD,HapM...
chr19	19296909	rs10401969	dbSNP		Hepatic lipid content in extreme obesity		1000Genomes,Cited,Frequency,gnomAD,HapMap,Pheno...
chr12	93554681	rs10859525	dbSNP		Hepatic lipid content in extreme obesity		1000Genomes,Cited,Frequency,gnomAD,HapMap,Pheno...
chr5	96391960	rs17085675	dbSNP	T	Monogenic Non-Syndromic Obesity	likely benign	1000Genomes,Cited,Frequency,gnomAD,HapMap,Pheno...
chr7	128254539	rs17151919	dbSNP	A	Monogenic Non-Syndromic Obesity	benign,likely benign	1000Genomes,Cited,ESPExAC,Frequency,gnomAD,HapM...
chr7	128255163	rs17151922	dbSNP	T	Monogenic Non-Syndromic Obesity	likely benign	1000Genomes,Cited,Frequency,gnomAD,HapMap,Pheno...
chr7	128256816	rs17617757	dbSNP	A	Monogenic Non-Syndromic Obesity	uncertain significance	1000Genomes,Frequency,gnomAD,Phenotype_or_Diseas...
chr1	65636595	rs34130975	dbSNP	C	Monogenic Non-Syndromic Obesity	uncertain significance	1000Genomes,ESPExAC,Frequency,gnomAD,Phenotype_...

Figure 3: VarPhen output

Customised list of genes

If instead of an OMIM id you are interested in a specific list of genes, you can use the *vargen_custom* function with a list of ensembl gene ids. This take the same input as vargen_pipeline (except the omim id) and follow the same workflow (variants on genes, promoters, gtex and gwas).

Example:

```
adipose_tissues <- select_gtex_tissues("./vargen_data/GTEx_Analysis_v8_eQTL/",
                                       "adipose")

obesity_traits <- list_gwas_traits("obesity", "./vargen_data/")

obesity_custom <- vargen_custom(vargen_dir = "./vargen_data/",
                                gene_ids = c("ENSG00000155846", "ENSG00000115138"),
```

```
                                outdir = "./",
                                gtex_tissues = adipose_tissues,
                                gwas_traits = obesity_traits,
                                verbose = T)
```

Tips

How to plot the gwas variants

If you want to visualise the variants in a manhattan plot, you can use the *plot_manhattan_gwas* function:

```
gwas_cat <- create_gwas("./vargen_data/")

alzheimer_traits <- c("Alzheimer's disease",
        "Alzheimer's disease (late onset)",
        "Alzheimer's disease biomarkers",
                "Alzheimer's disease (cognitive decline)")

plot_manhattan_gwas(gwas_cat = gwas_cat, traits = alzheimer_traits)

# Optional: if you want to save the plot as a pdf
grDevices::dev.print(pdf, "./manhanttan_alzheimer.pdf")
```

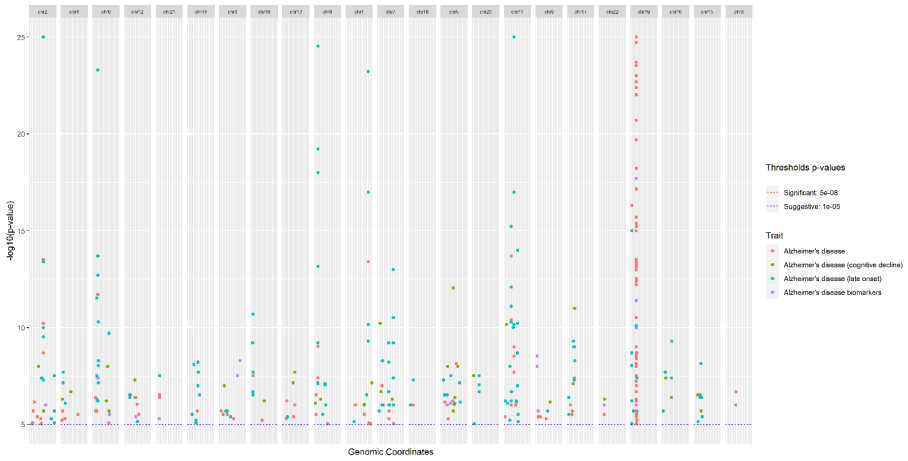


Figure 4: Example of manhanttan plot for alzheimer's disease

The two thresholds, suggestive and significant, correspond to the definition given by Lander and Kruglyak:

```
Lander E, Kruglyak L. Genetic dissection of complex traits:
guidelines for interpreting and reporting linkage results. Nat Genet. 1995;11(3):241-7.
```

How to plot the omim variants

You can use the *vargen_visualisation* function to have an overview of the variants located on the omim genes. Note that one variant can be represented multiple times, as it consequence and phred score will be different according to each transcript.

```
# cf: 'Annotating the variants' to create "vargen_variants_annotated.tsv":
obesity_vargen_ann <- read.table(file = "./OMIM_601665/vargen_variants_annotated.tsv",
                                 header = T, sep = "\t", stringsAsFactors = F)

gene_mart <- connect_to_gene_ensembl()
vargen_visualisation(annotated_snps = obesity_vargen_ann,
                     outdir = "./obesity_gviz/",
                     device = "png",
                     gene_mart = gene_mart)
```

The plot contains 4 tracks: - The chromosome, with a red marker on the gene location - The ensembl transcripts - The variant consequences, grouped by type (eg: INTRONIC, STOP LOST etc...), each green bar represent a variant - The cadd phred score, each dot represent a variant.

Figure 5: vargen visualisation

The **rsid_highlight** parameter allows you to highlight some variants (by rsid) in red:

```
obesity_vargen_ann <- read.table("OMIM_601665/vargen_variants_annotated.tsv")
```

```
gene_mart <- connect_to_gene_ensembl()
vargen_visualisation(annotated_snps = obesity_vargen_ann, verbose = T,
                     outdir = "./obesity_gviz/", device = "png",
                     gene_mart = gene_mart,
                     rsid_highlight =
        unique(
            obesity_vargen_ann
            [obesity_vargen_ann$cadd_phred > 20, "rsid"]))
```

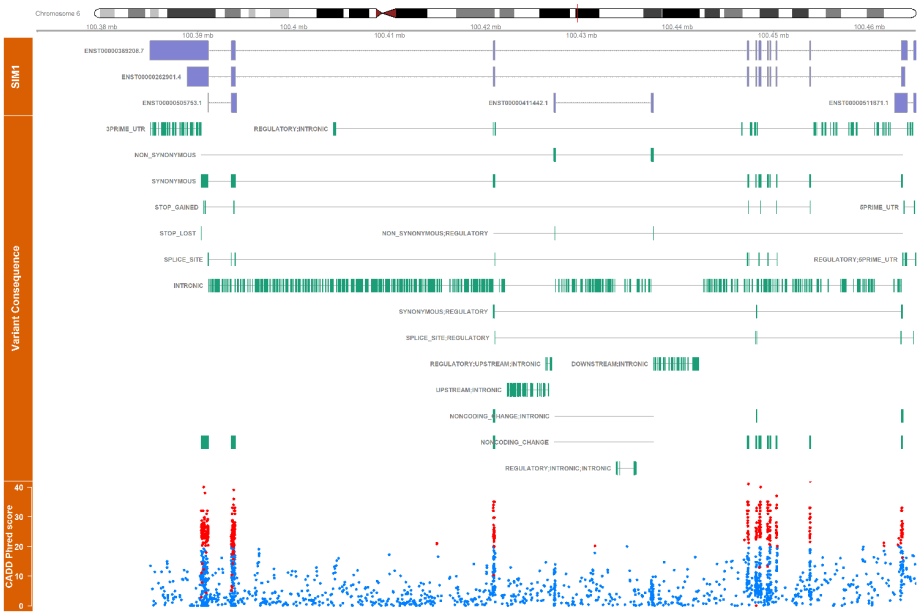


Figure 6: vargen visualisation with highlighted variants

More examples

In the following example we are assuming that the files described in "Installation" have been installed in the folder "./vargen_data", either manually or by running:

```
vargen_install("./vargen_data")
```

- Example 1: "Simple query for Type 1 Diabetes Mellitus"

```
T1DM_variants <- vargen_pipeline(vargen_dir = "./vargen_data",
                                 omim_morbid = "222100")
```

- Example 2: "Type 1 Diabetes Mellitus with GTEx"

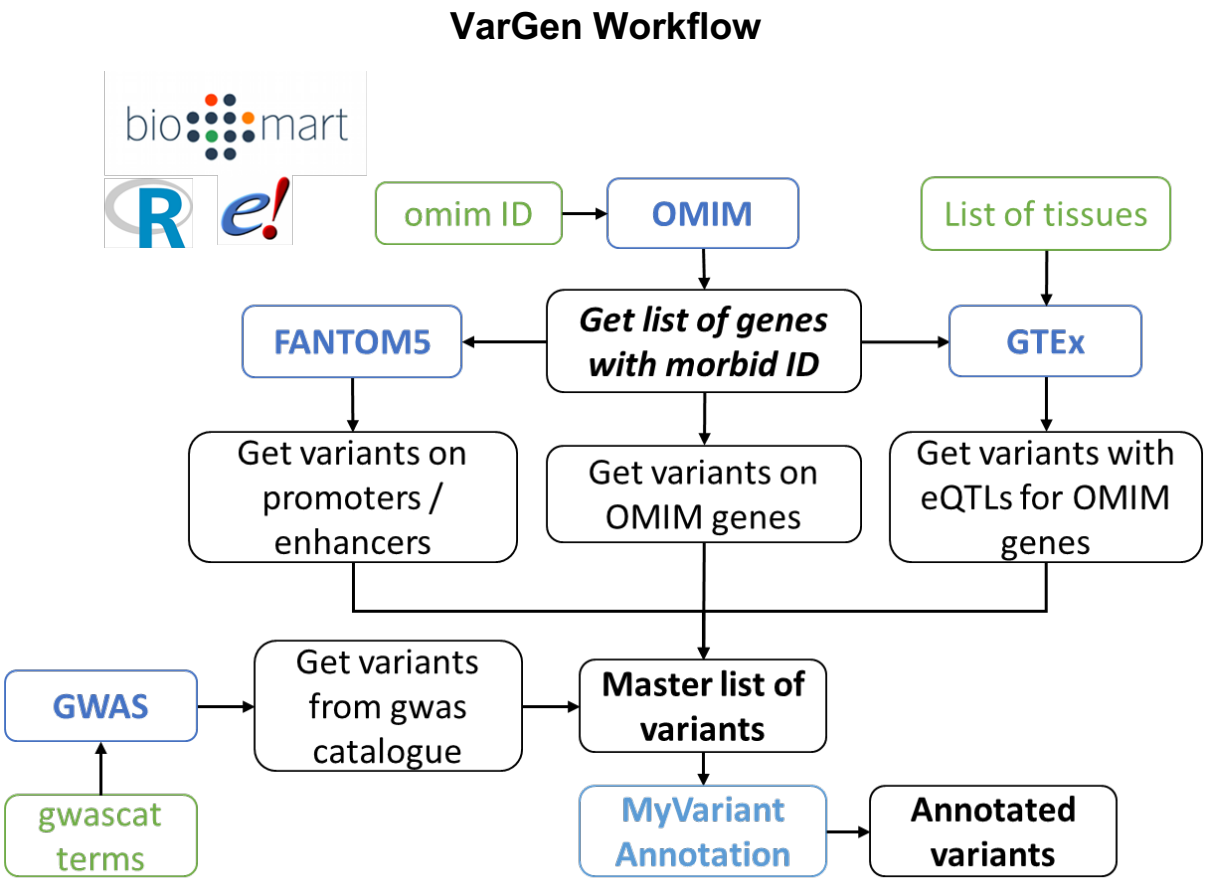
```
# First select the tissues of interest (here pancreas)
# it is possible to select more than one tissue (eg: c("pancreas", "adipose"))
```

```
pancreas_gtex <- select_gtex_tissues(gtex_dir = "./vargen_data/GTEx_Analysis_v8_eQTL/",
                                     tissues_query = "pancreas")

T1DM_variants <- vargen_pipeline(vargen_dir = "./vargen_data",
                                 omim_morbid = "222100",
                                 gtex_tissues = pancreas_gtex)
```

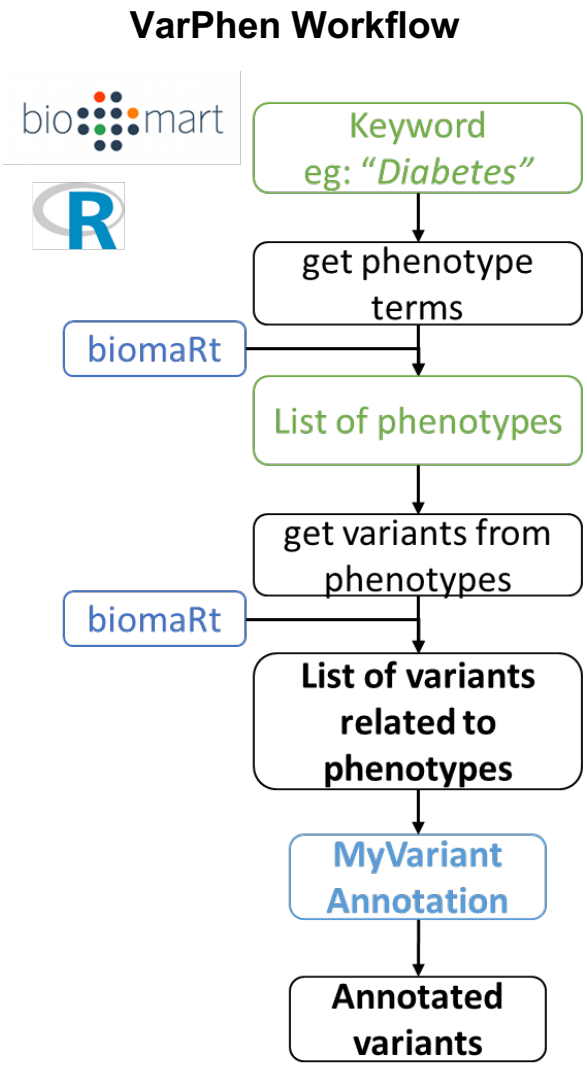
- Example 3: "Obesity with GTEx, GWAS and annotation"

```
# First select the tissues of interest (here adipose)
# it is possible to select more than one tissue (eg: c("pancreas", "adipose"))
adipose_gtex <- select_gtex_tissues(gtex_dir = "./vargen_data/GTEx_Analysis_v8_eQTL/",
                                    tissues_query = "adipose")

# List the available gwas traits:
list_gwas_traits(keywords = "obesity")

# Select gwas traits of interest
gwas_obesity <- c("Obesity (extreme)",
        "Obesity-related traits",
        "Obesity", "Obesity (early onset extreme)")

Obesity_variants <- vargen_pipeline(vargen_dir = "./vargen_data",
                                    omim_morbid = "601665",
                                    gtex_tissues = adipose_gtex,
                                    gwas_traits = gwas_obesity,
                                    fantom_corr = 0.25,
                                    verbose = TRUE)

# Annotation of the variants obtained with the previous command:
Obesity_annotation <- annotate_variants(rsid = Obesity_variants$rsid)

# Merging the original output with the annotation:
Obesity_variants_annotated <- merge(Obesity_variants, Obesity_annotation)
View(Obesity_variants_annotated)
```

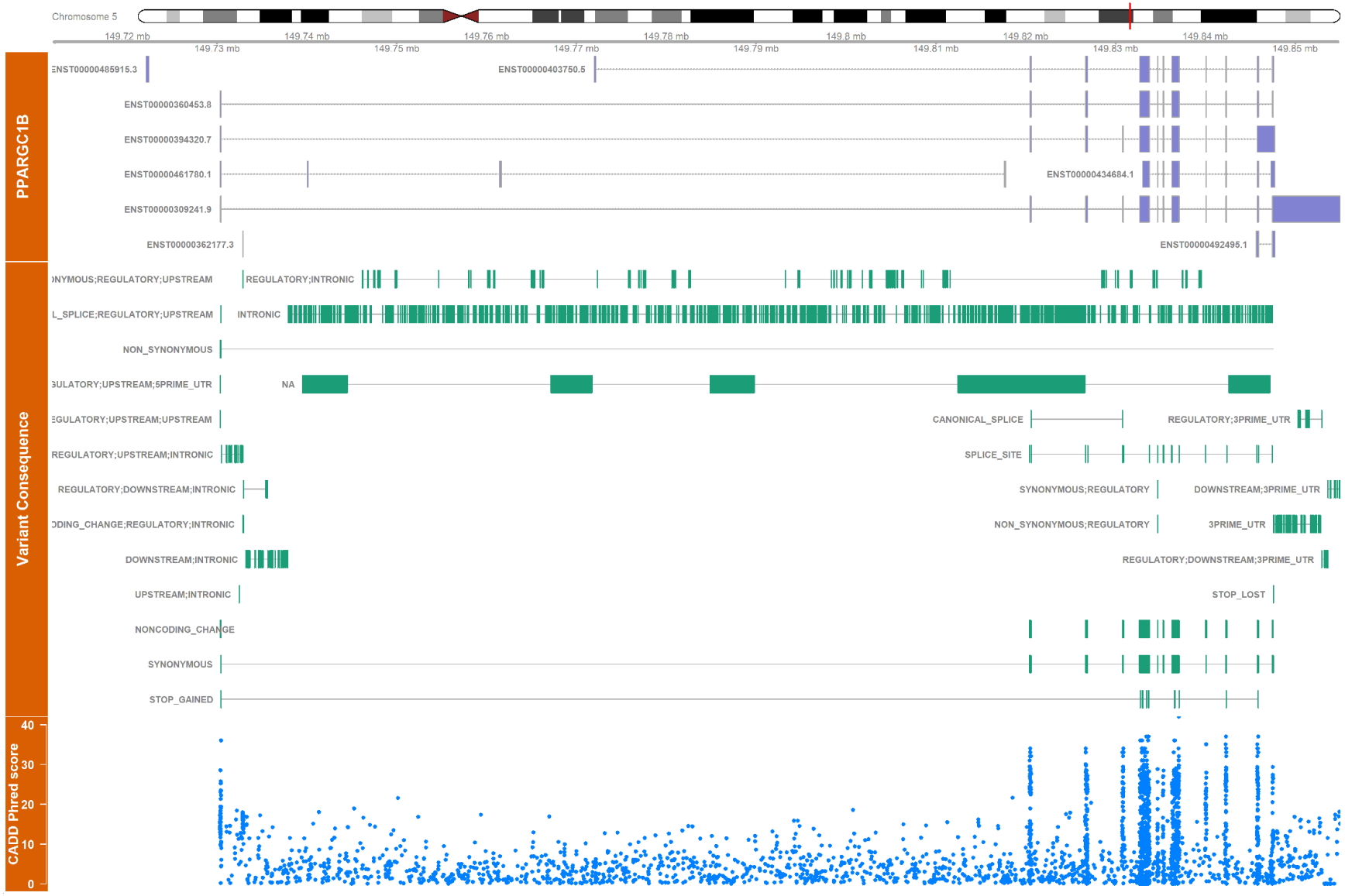
Supplementary Material S2 - Supplementary Figures:



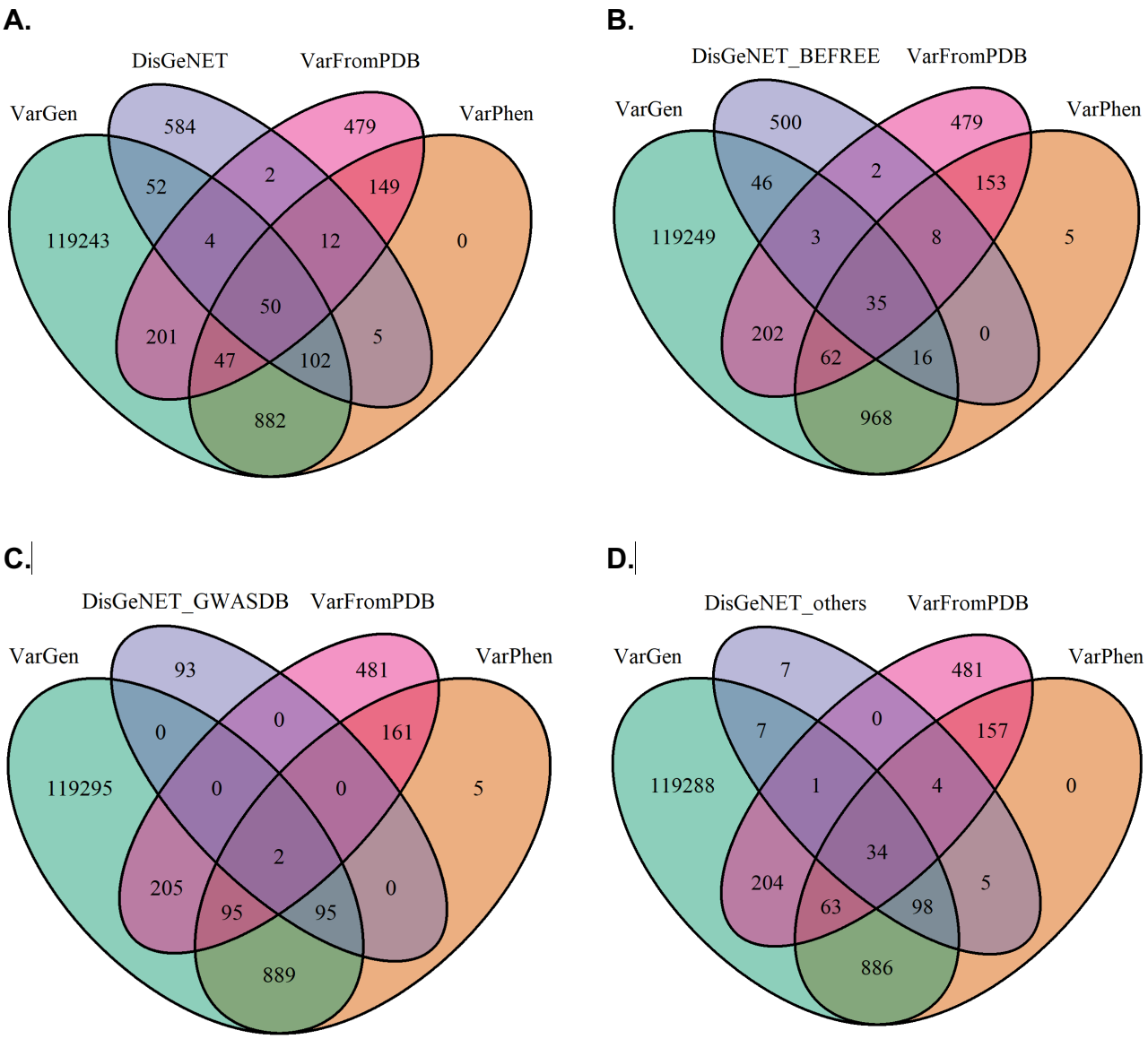
Supplementary Figure 1: VarGen workflow: user inputs and public databases are represented by green and blue boxes respectively. The pipeline is centred on the list of genes related to the disease in the Online Mendelian Inheritance in Man (subsequently called the "OMIM genes"). VarGen gets the variants located directly on the OMIM genes, as well as on their enhancers and promoters, using FANTOM5. GTEx is used to get tissue-specific variants impacting the expression of the OMIM genes. Finally, the user can also input a list of gwas traits to get variants from gwas catalogue. The package also offers the possibility to annotate the variants. The databases are accessed using BiomaRt, Ensembl API and local files. The only mandatory input is an OMIM ID or alternatively, a customised list of genes.



Supplementary Figure 2: VarPhen workflow: user inputs and public databases are represented by green and blue boxes respectively. VarPhen is a more specific, alternative pipeline, available within the package. The pipeline starts by fetching biomaRt to get all the phenotypes related to a user entered keyword. The user can then use one or more of these phenotypes to get the list of associated variants in biomaRt. As with VarGen it is possible to annotate them to obtain further information the impact of each variant.

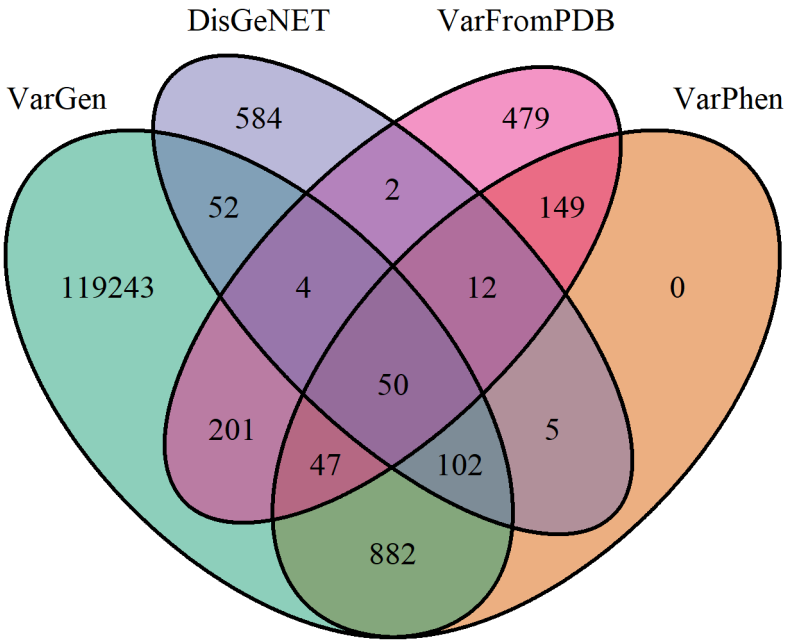


*Supplementary Figure 3: Visualisation created with VarGen for the PPARGC1B gene. In order to have an overview of the variants, the **vargen_visualisation** function was written. This plot is created for each gene and contains four tracks, from top to bottom: 1) the chromosome with a red marker on the gene locus 2) the ensembl transcripts with the exons represented as purple blocks 3) the variants represented by green bars and grouped by consequences 4) The variants represented by blue dots and plotted according to their cadd phred score, with higher scores meaning a more deleterious variant. It is also possible to highlight some variants by rsid (not showed in the figure).*

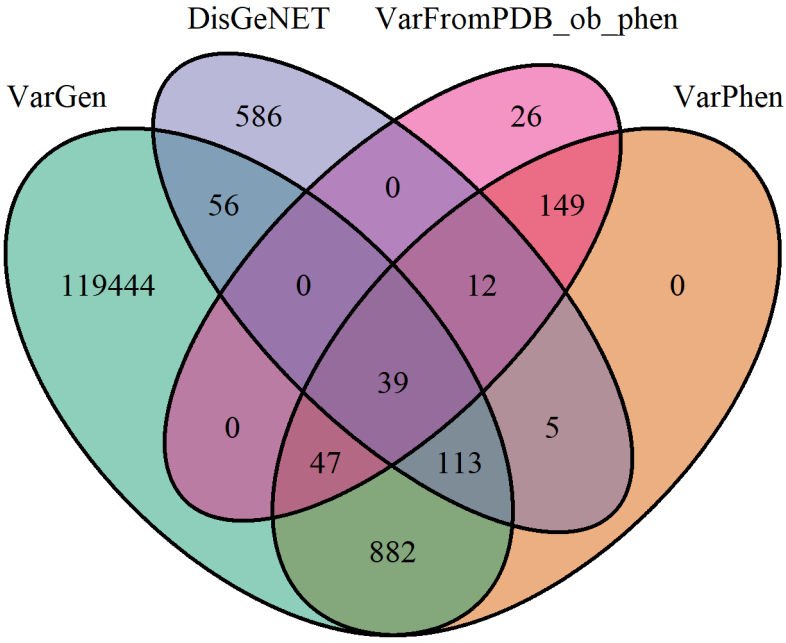


Supplementary Figure 4: Venn diagrams representing the variants found by the different pipelines during benchmarking: VarGen, DisGeNET, V and VarPhen. Obesity (OMIM: 601665) was chosen as the use case. ***A.*** *Results with the raw output for all the pipelines.* ***B.*** *Results using only the literature mining (BEFREE) for DisGeNET.* ***C.*** *Results using only the gwasDB database for DisGeNET* ***D.*** *Results using the others DisGeNET databases (gwasdb, uniprot and clinvar). Most of the variants unique to DisGeNET are coming from the literature mining and gwasDB.*

A.

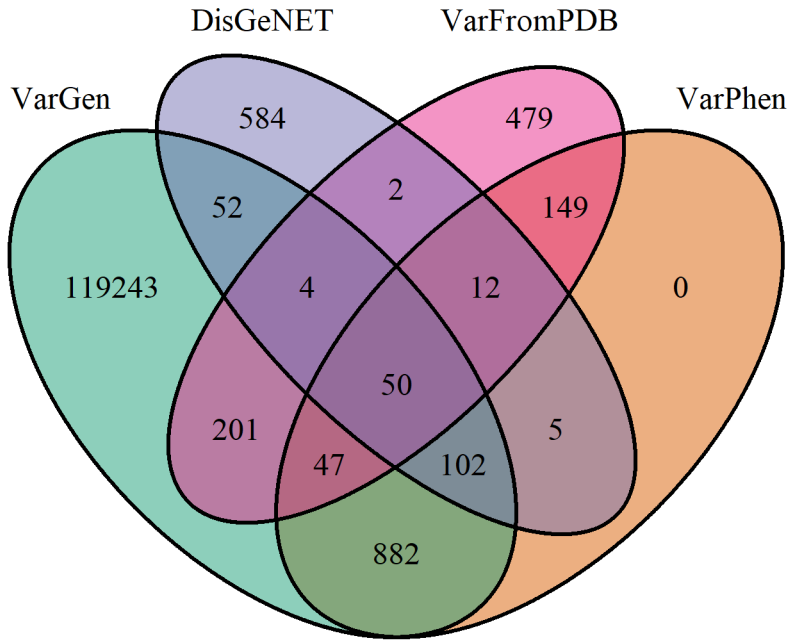


B.

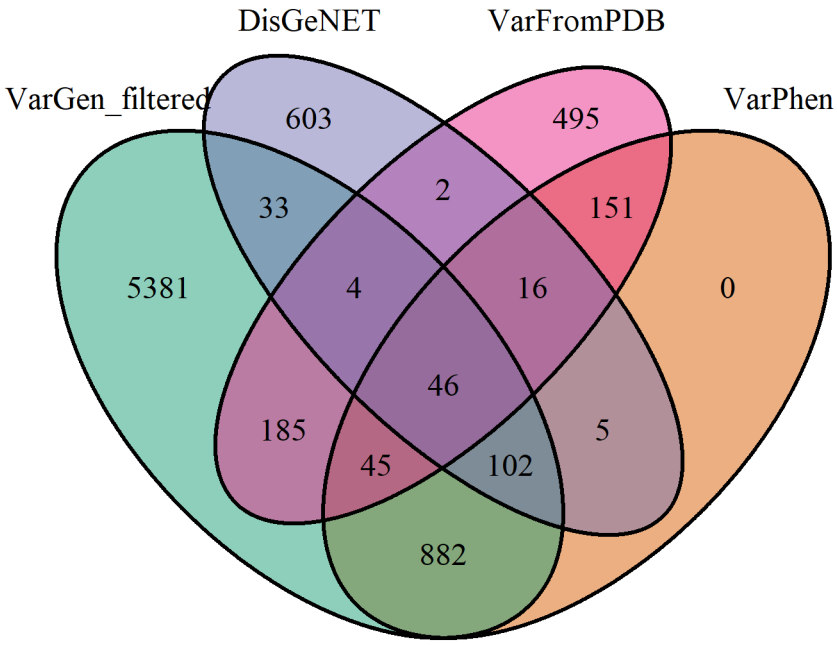


*Supplementary Figure 5: Venn diagrams representing the variants found by the different pipelines during benchmarking: VarGen, DisGeNET, VarFromPDB and VarPhen. Obesity (OMIM: 601665) was chosen as the use case**. A.** Results with the raw output for all the pipelines. **B**. Results obtained when filtering the variants by phenotypes containing the word "obesity" for VarFromPDB. Most of VarFromPDB unique variants are not coming directly from phenotypes containing the word "obesity".*

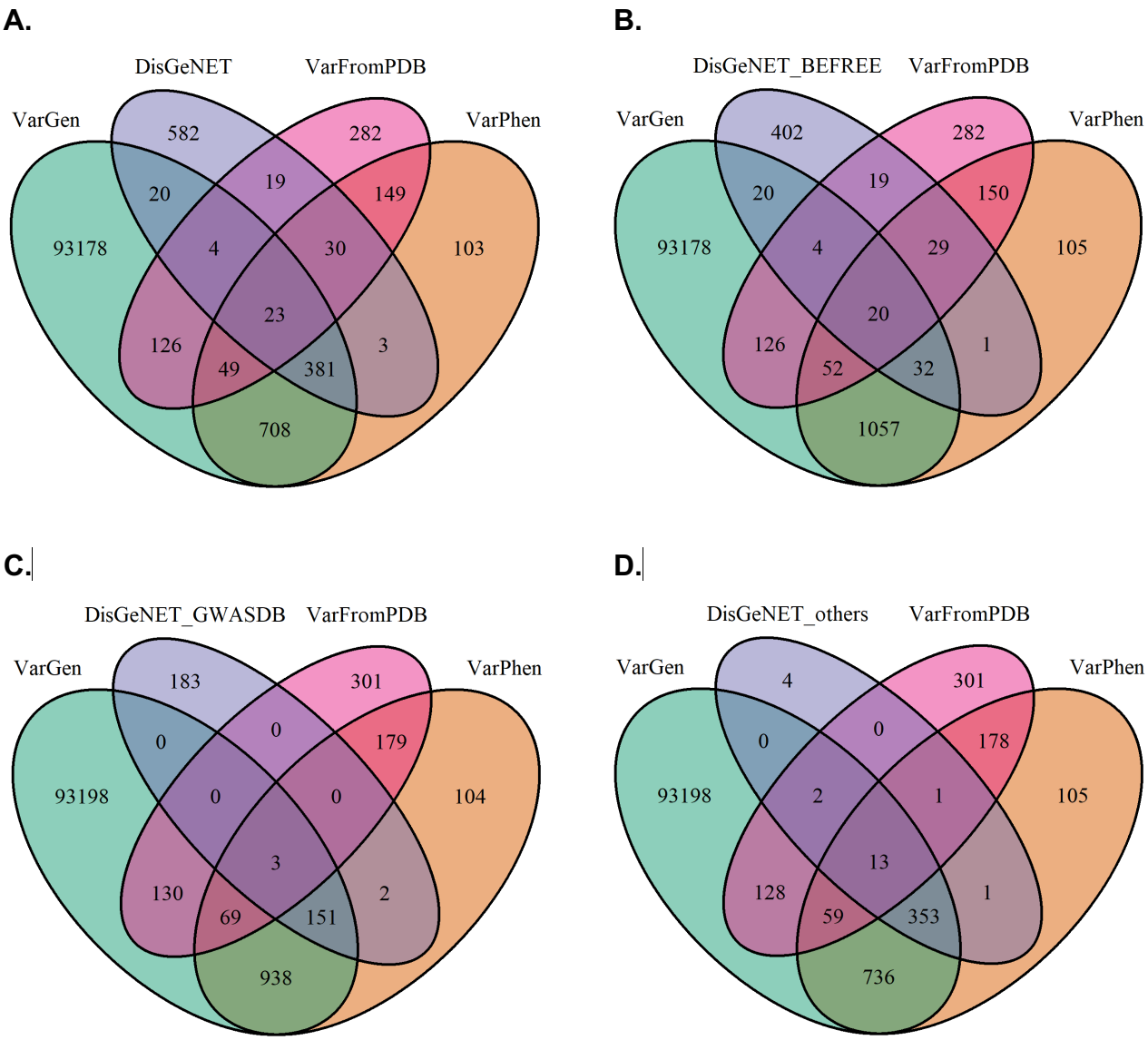
A.



B.

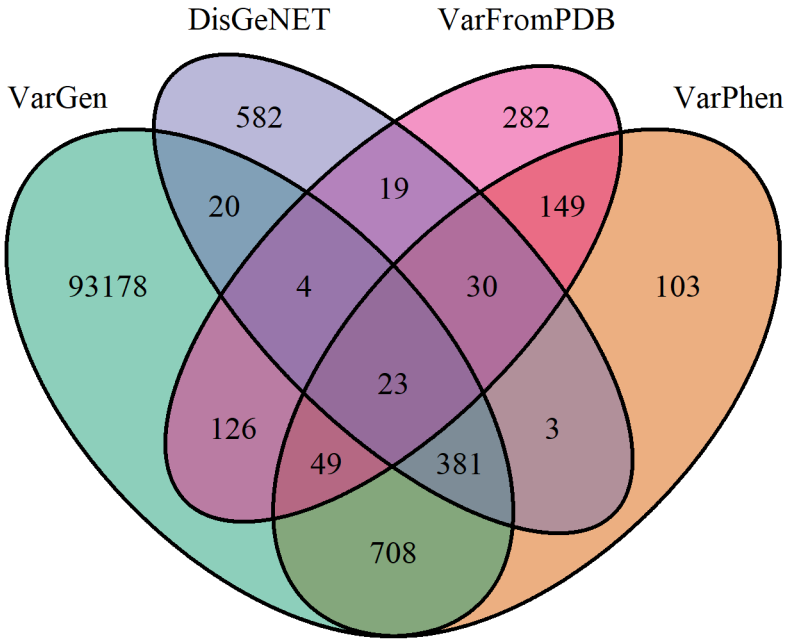


Supplementary Figure 6: Venn diagrams representing the variants found by the different pipelines during benchmarking:,VarGen, DisGeNET, VarFromPDB and VarPhen. Obesity (OMIM: 601665) was chosen as the use case. ***A.*** *Venn diagram using the raw output for all the pipelines.* ***B.*** *Venn diagram using the filtered VarGen dataset, with the following strategy: all the variants from the gwas catalogue and with clinical significance were kept, and the rest were filtered if their cadd phred score was below 10.*

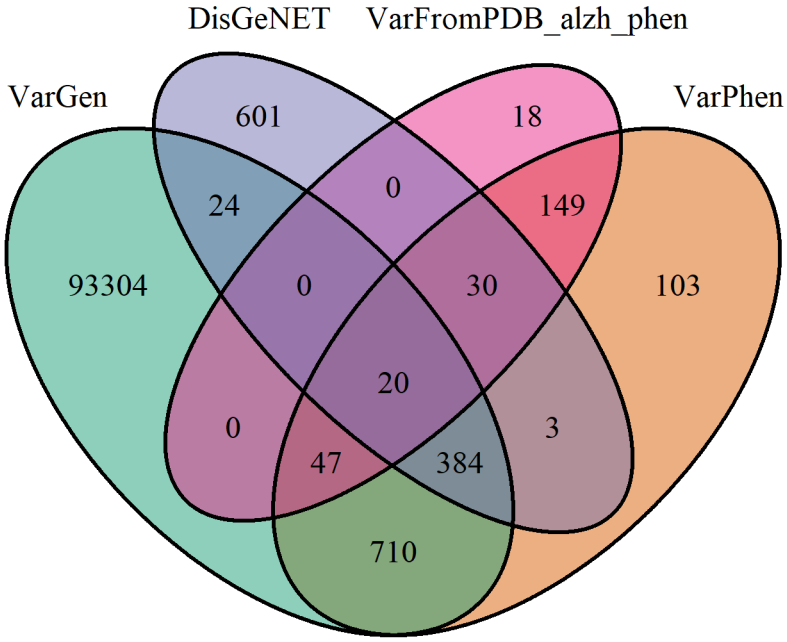


Supplementary Figure 7: Venn diagrams representing the variants found by the different pipelines during benchmarking: VarGen, DisGeNET, V and VarPhen. Alzheimer (OMIM: 104300) was chosen as the use case. ***A.*** *Results with the raw output for all the pipelines.* ***B.*** *Results using only the literature mining (BEFREE) for DisGeNET.* ***C.*** *Results using only the gwasDB database for DisGeNET* ***D.*** *Results using the others DisGeNET databases (gwasdb, uniprot and clinvar). Most of the variants unique to DisGeNET are coming from the literature mining and gwasDB.*

A.

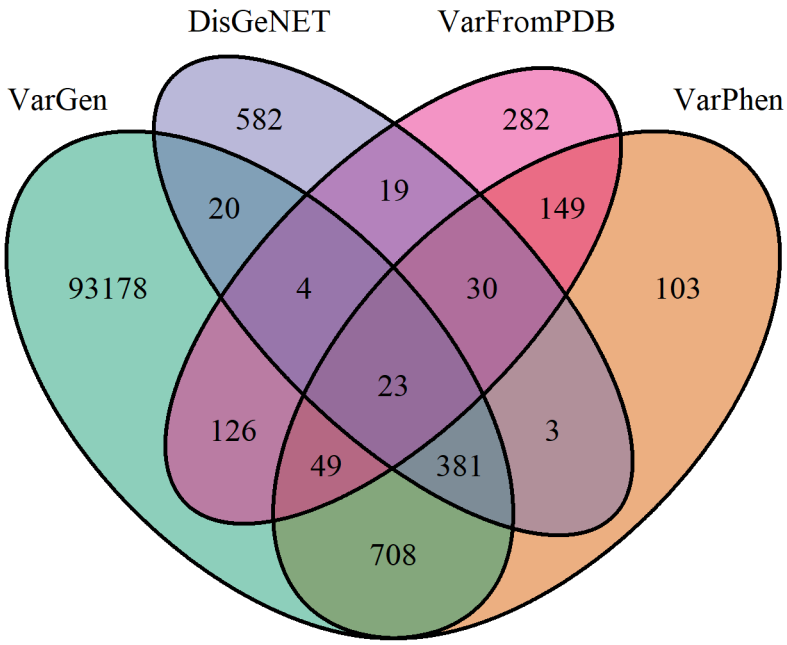


B.

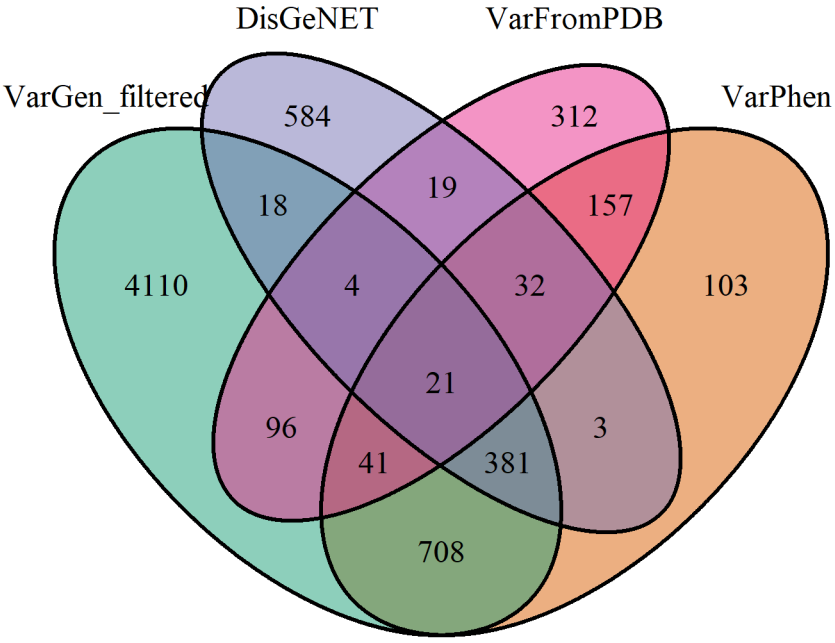


*Supplementary Figure 8: Venn diagrams representing the variants found by the different pipelines during benchmarking: VarGen, DisGeNET, VarFromPDB and VarPhen. Alzheimer (OMIM: 104300) was chosen as the use case. **A.** Results with the raw output for all the pipelines. **B**. Results obtained when filtering the variants by phenotypes containing the word "alzheimer" for VarFromPDB. Most of VarFromPDB unique variants are not coming directly from phenotypes containing the word "alzheimer".*

A.



B.



Supplementary Figure 9: Venn diagrams representing the variants found by the different pipelines during benchmarking:,VarGen, DisGeNET, VarFromPDB and VarPhen. Alzheimer (OMIM: 104300) was chosen as the use case. ***A.*** *Venn diagram using the raw output for all the pipelines.* ***B.*** *Venn diagram using the filtered VarGen dataset, with the following strategy: all the variants from the gwas catalogue and with clinical significance were kept, and the rest were filtered if their cadd phred score was below 10.*

Supplementary Tables:

Supplementary Table 1: List of phenotypes found with VarFromPDB using the keyword "Obesity" as input

not provided	Progressive sclerosing poliodystrophy
not specified	OBESITY (BMIQ14), SUSCEPTIBILITY TO
Arterial calcification of infancy	Autistic behavior
Hypophosphatemic Rickets, Recessive	Myopia
Abetalipoproteinaemia	Seizures
Inborn genetic diseases	Strabismus
Abnormality of neuronal migration	Coffin-Siris syndrome
Monogenic Non-Syndromic Obesity	Rhabdoid tumor predisposition syndrome 2
Proopiomelanocortin deficiency	intellectual deficiency
History of neurodevelopmental disorder	Hearing impairment
Intellectual Disability, Recessive	Hypothyroidism
Mental retardation, autosomal recessive 13	Intellectual disability
Joubert syndrome	Short metacarpal
Obesity, autosomal dominant	Short stature
Monogenic diabetes	Truncal obesity
Obesity	Attention deficit hyperactivity disorder
salbutamol response - Efficacy	Delayed speech and language development
salmeterol response - Efficacy	Infantile muscular hypotonia
antipsychotics response - Toxicity/ADR	Malar flattening
amisulpride response - Toxicity/ADR	Mandibular prognathia
aripiprazole response - Toxicity/ADR	Misalignment of teeth
clozapine response - Toxicity/ADR	Poor fine motor coordination
haloperidol response - Toxicity/ADR	Thin upper lip vermilion
olanzapine response - Toxicity/ADR	Brachydactyly
paliperidone response - Toxicity/ADR	Cognitive impairment
quetiapine response - Toxicity/ADR	Cushing's syndrome
risperidone response - Toxicity/ADR	Hypocalcemia
ziprasidone response - Toxicity/ADR	McCune-Albright syndrome
DEVELOPMENTAL DELAY, INTELLECTUAL DISABILITY, OBESITY, AND DYSMORPHISM	Short stature, brachydactyly, intellectual developmental disability, and seizures
EPILEPTIC ENCEPHALOPATHY, EARLY INFANTILE, 58	Progressive osseous heteroplasia
Leptin receptor deficiency	Pseudohypoparathyroidism
Diabetes Mellitus, Noninsulin-Dependent, with Acanthosis Nigricans and Hypertension	Pseudohypoparathyroidism type 1B
Familial partial lipodystrophy	Pseudohypoparathyroidism type 1C
Proprotein convertase 1/3 deficiency	Pseudopseudohypoparathyroidism
Leptin deficiency or dysfunction	Round face
Spastic paraplegia, intellectual disability, nystagmus, and obesity	Subcutaneous nodule
Joubert Syndrome and Related Disorders	Tetany
Obesity, hyperphagia, and developmental delay	Abnormal facial shape

Bardet-Biedl syndrome	Abnormality of the dentition
Syndromic intellectual disability	Acanthosis nigricans
Familial partial lipodystrophy 3	Hepatic steatosis
BODY MASS INDEX QUANTITATIVE TRAIT LOCUS 20	Hyperlipidemia
Neuroblastoma	Insulin resistance
Diabetes mellitus type 2	Intellectual disability, severe
Hypophosphatemic rickets, autosomal recessive, 2	Lumbar hyperlordosis
PHIP-Related disorders	Renal hypoplasia
Schizophrenia	Self-injurious behavior
-	PITUITARY ADENOMA 3, MULTIPLE TYPES
LEPR-Related Disorders	Abnormal platelet function
POMC-Related Disorders	Abnormal platelet morphology
ENPP1-Related Disorders	Abnormal platelet shape
Joubert syndrome 1	Epistaxis
MORM syndrome	Hypertension
Absent epiphyses	Increased mean platelet volume
Chronic lung disease	Numerous pigmented freckles
Cleft palate	Thrombocytopenia
Coat hanger sign of ribs	Difficulty walking
Hemivertebrae	Generalized hypotonia
Interstitial pulmonary abnormality	Gynecomastia
Micrognathia	Horizontal nystagmus
Patent ductus arteriosus	Muscular hypotonia
Preaxial foot polydactyly	Pes planus
Pseudoarthrosis	Poor motor coordination
Respiratory failure	Sleep disturbance
Short femur	Neurological speech impairment
Skeletal dysplasia	Pigmentary retinopathy
Talipes equinovarus	Immunodeficiency
Vertebral hypoplasia	Protruding tongue
Vertebral segmentation defect	Severe T-cell immunodeficiency
Metabolic syndrome, susceptibility to	Cardiomyopathy
Obesity, age at onset of	Gastroesophageal reflux
Obesity, mild, early-onset	Glaucoma
Leanness, inherited	Limb-girdle muscle weakness
Obesity, late-onset	Muscle weakness
Obesity, severe, and type II diabetes	Obstructive sleep apnea syndrome
UCP3 POLYMORPHISM G/A	Panhypopituitarism
Morbid obesity	Specific learning disability
Body mass index, modifier of	Microcephaly
Diabetes mellitus, noninsulin-dependent, modifier of	Obesity, variation in
Intimal medial thickness of internal carotid artery, modifier of	Bardet-Biedl syndrome 8
Obesity, modifier of	Intellectual disability, moderate
Carcinoma of colon	Postaxial foot polydactyly
Glioma susceptibility 1	Non-ketotic hyperglycinemia

LEPTIN RECEPTOR POLYMORPHISM	Cardiovascular phenotype
Obesity, association with	Congenital long QT syndrome
Obesity, early-onset, susceptibility to	Long QT syndrome
Insulin resistance, susceptibility to	Long QT syndrome 2
Body mass index quantitative trait locus 12	Prolonged QT interval
Metabolic syndrome, protection against	Asthma
Asthma, nocturnal, susceptibility to	Breast-ovarian cancer, familial 2
ADRB2 POLYMORPHISM	Ectopic ossification
Beta-2-adrenoreceptor agonist, reduced response to	Headache
Body mass index quantitative trait locus 18	Hereditary breast and ovarian cancer syndrome
Cole disease	Hereditary cancer-predisposing syndrome
Morbid obesity and spermatogenic failure	Migraine
Abdominal obesity-metabolic syndrome 3	Nephrolithiasis
See cases	Short attention span
Leptin dysfunction	Striae distensae
Retinal dystrophy and obesity	Spastic paraplegia 54, autosomal recessive
PHIP-Related Disorder	Dolichocephaly
Generalized epilepsy	Macrocephalus
Global developmental delay	NA
Mild obesity	

Supplementary Table 2: List of phenotypes found with VarFromPDB using the keyword "Alzheimer" as input

Alzheimer disease, susceptibility to	Genetic prion diseases
Cutaneous photosensitivity	Gerstmann-Straussler-Scheinker syndrome
Hemochromatosis type 1	Huntington disease-like 1
Hemochromatosis type 2	Jakob-Creutzfeldt disease
Hemochromatosis, juvenile, digenic	Inborn genetic diseases
Hereditary cancer-predisposing syndrome	Alzheimer disease, early-onset, susceptibility to
Hereditary hemochromatosis	Aphasia, primary progressive, susceptibility to
Microvascular complications of diabetes 7	Fatal familial insomnia
Porphyria cutanea tarda, susceptibility to	Prion disease, susceptibility to
Porphyria variegata, susceptibility to	not specified
Porphyrinuria	Spongiform encephalopathy with europsychiatric features
Transferrin serum level quantitative trait locus 2	Protection against Creutzfeldt-Jakob disease
not provided	Kuru, protection against
Alzheimer's disease	Coronary artery spasm 1, susceptibility to
Familial porphyria cutanea tarda	Hypertension resistant to conventional therapy
Variegate porphyria	Hypertension, pregnancy-induced, susceptibility to
HFE INTRONIC POLYMORPHISM	Ischemic heart disease, susceptibility to
HFE POLYMORPHISM	Ischemic stroke, susceptibility to
Myeloperoxidase deficiency	Metabolic syndrome, susceptibility to
POLYCYSTIC LIPOMEMBRANOUS OSTEODYSPLASIA WITH SCLEROSING LEUKOENCEPHALOPATHY 2	Apolipoproteinemia E1
Polycystic lipomembranous osteodysplasia with sclerosing leukoencephalopathy	Familial type 3 hyperlipoproteinemia
Alzheimer disease, type 4	Warfarin response
Dilated cardiomyopathy 1V	atorvastatin response - Efficacy
Alzheimer disease, late-onset, susceptibility to	HYPERLIPOPROTEINEMIA, TYPE III, AND ATHEROSCLEROSIS ASSOCIATED WITH APOE5
Alzheimer disease, familial, 3, with unusual plaques	HYPERLIPOPROTEINEMIA, TYPE III, AND ATHEROSCLEROSIS ASSOCIATED WITH APOE7
Dilated cardiomyopathy	APOE2-DUNEDIN
Heart failure	APOE5 VARIANT
ALPHA-2-MACROGLOBULIN POLYMORPHISM	APOE4 VARIANT
CEREBRAL AMYLOID ANGIOPATHY, PRNP-RELATED	APOE4(-)-FREIBURG
Quebec platelet disorder	Alzheimer disease 2
Alzheimer disease, protection against	Primary degenerative dementia of the Alzheimer type, presenile onset
Major depressive disorder	APOE3(-)-FREIBURG

Reticulate acropigmentation of Kitamura	APOE3 VARIANT
Alzheimer disease 18	APOE2 VARIANT
Dilated Cardiomyopathy, Dominant	APOE4(+)
Early-Onset Familial Alzheimer Disease	Myocardial infarction
Dementia	Lipoprotein glomerulopathy
Mental deterioration	Cerebral amyloid angiopathy, APP-related
Alzheimer disease familial 3, with spastic paraparesis	Alzheimer disease, type 1
Sea-blue histiocyte syndrome	CEREBRAL AMYLOID ANGIOPATHY, APP-RELATED, PIEDMONT VARIANT
See cases	Alzheimer disease, type 3
Primary dilated cardiomyopathy	Acne inversa, familial, 3
Hirschsprung disease 1	Frontotemporal dementia
Alzheimer disease, type 9	Pick's disease
APP POLYMORPHISM	Alzheimer disease, familial, with spastic paraparesis and unusual plaques
Abnormality of lipid metabolism	Cardiomyopathy, dilated, 1u
Huntington disease-like syndrome	Alzheimer disease, familial, 3, with spastic paraparesis and apraxia
Hereditary cerebral hemorrhage with amyloidosis	PRNP-associated condition
Vascular dementia, susceptibility to	Leigh syndrome
Atransferrinemia	Parkinson disease, late-onset
Transferrin variant c1/c2	Alzheimer disease 19
NA	